 NOT TO BE FURTHER COPIED OR REPRODUCED

(* y VS. x *)

9

Straightening Out Plots
(using three points)

Introduction 9-1

9A. Looking at three points 9-2

9B. Re-expressing y alone 9-4
- U. S. population again, 9-6

9C. Re-expressing x alone 9-7
- back to the population of the U.S., 9-8
- a caveat, 9-12

9D. A braking example 9-16
- using our knowledge, 9-21

9E. The vapor pressure of H_2O 9-25

9F. Re-expressing both x and y 9-32
- another try, 9-35

9G. Wise change of origin as a preliminary 9-37
- a radioactive decay example, 9-38

* * *

9H. How far have we come? 9-45

* * *

9P. Problems 9-47

Exh.		Page
	A	
	B	
	C	
1		9-9
2		9-13
3		9-14
4		9-15
	D	
5		9-17
6		9-18
7		9-20
8		9-22
9		9-23
10		9-24
11		9-26
12		9-27
	E	
13		9-29
14		9-30
15		9-31
16		9-33
	F	
	G	
17		9-39
18		9-40
19		9-42
20		9-43
	H	
21		9-46
	P	
22		9-48
23		9-49

9

Straightening Out Plots

(using three points)

We are now sure that we want to first straighten out and then flatten out plots. Straightening out is important -- we want to learn to do it as easily as we reasonably can. This chapter is devoted to techniques and examples.

The sort of re-expression that concerns us is -- as we have said -- almost entirely re-expression of amounts and large counts. These are the kinds of numbers where it is natural to take powers, roots, and logs.

In dealing with them, we will want to be sure that our origins are reasonably chosen. It is only for amounts measured from reasonable origins that we are likely to get full value from changing to a power, a root or a logarithm.

Powers and roots of amounts are again amounts. Logs of amounts are balances. For these purposes -- as for so many others -- large counts are merely a special kind of amount. Accordingly, powers and roots of large counts are amounts, while logs of large counts are balances.

In thinking about our problems of re-expression -- which need not be the same sort of thing as analyzing the data involved -- we need to think about whether x -- or y -- varies much or little. So long as we deal with amounts,

the natural way to make comparisons is by ratios -- including %s. Thus we are interested in such facts as

$$\frac{\text{largest } x}{\text{smallest } x} = 3$$

all x are within ±30% of a middle value

$$\frac{\text{largest } x}{\text{smallest } x} = 1.1$$

largest x = smallest x plus 10%.

When we work in logs, we are concerned with balances, not amounts, and the natural way of expressing spread is either in log units or -- occasionally -- by the ratios into which differences in log units can be re-expressed.

9A. Looking at three points

We have seen various illustrations of straightening out data. So far, either these have some from simple rational considerations, or they have come -- "out of the air" -- with little apparent reason for the choice. How are we to behave when we have some other kind of data? Do we have to try all possible combinations of an expression of y with an expression of x on the whole data? Or can we save most of the effort this would involve?

The purpose of these changes of expression is to straighten out the data. If the data looks curved, in some overall way, we can make this obvious by picking out three representative points. For the early growth of the U.S.A. population, for example, we could choose the points corresponding to 1800, 1850, and 1890. These three points are:

(1800, 5.3)
(1850, 23.2)
(1890, 62.9)

A simple way to see that three points do not lie on a single straight line is to find the slopes of the straight lines through the first pair and the second pair, respectively, this gives:

$$\frac{23.2-5.3}{1850-1800} = \frac{17.9}{50} = 0.36$$

and

$$\frac{62.9-23.2}{1890-1850} = \frac{39.7}{40} = 0.99$$

which are quite different. The second slope is greater, so the curve is hollow upward. (Draw yourself a sketch.)

If any pair of choices of expression is going to straighten out the early portion of the U.S.A. population curve, these same choices will have to do a reasonably good job in straightening out these three points. We can save a lot of effort by screening our pairs of expressions on these three points. We will then need to try only the one or few best on the whole data.

We can often ease the task a little by choosing the spacing of the three points to simplify matters further. Had we chosen

(1810, 7.2)
(1850, 23.2)
(1890, 62.9)

We could have compared

$$16.0 = 23.2 - 7.2$$

with

$$39.7 = 62.9 - 23.2$$

since both changes in x are the same. (1850 - 1810 = 40 = 1890 - 1850).

Sometimes we can go almost this far by picking points so that changes in x have a simple ratio -- 1 to 2, 1 to 3, 3 to 2, etc. -- instead of being equal.

Even trying all reasonably possible pairs of expressions on three points is an effort. Can we save some of this by looking into the direction in which changing an expression shifts curvature?

9B. Re-expressing y alone

Exhibit 19 of chapter 3 shows various expressions of y plotted against y. We see at once that the higher curves are hollow upward, while the lower curves are hollow downward.

To say that a curve is hollow upward means that if we take 3 points on the curve, the middle point is below the line joining the other two. Similarly, hollow downward means that the middle point is above the line joining the other two. Since we are trying to get a middle point onto the line joining the outer two, these are just the sort of facts that matter to us.

We can say more about the simple ladder of ways of expression which includes

$$y^3$$

$$y^2$$

$$y$$

$$\sqrt{y}$$

$$\log y$$

$$-\frac{1}{\sqrt{y}}$$

$$-\frac{1}{y}$$

$$-\frac{1}{y^2}$$

$$-\frac{1}{y^3}$$

than we just have. We stated one special case of the following:

- If one expression is straight, those above it are hollow upward, those below it are hollow downward.

Exhibit 19 of chapter 3 already showed us that this is true when y is straight. Exhibit 20 of chapter 3 shows us that this is true when log y is straight. The appearance of both those exhibits makes it plausible that the statement is true if any expression of the simple ladder is straight. (The doubting reader should try to use these two statements to construct a general proof.)

- if our three points are hollow upward, we look further down the ladder for straightness.

- if our three points are hollow downward, we look further up the ladder for straightness.

These must be true. Consider the first: If a new expression is to be straight, and our present expression is hollow upward, the present expression has to be higher up the ladder than the new expression. To find the new expression we must move down the ladder from the present expression.

So far as re-expressing y goes, the rule is simple:

- move on the ladder as the bulging side of the curve points.

(* U.S. population again *)

Applying this to the early U.S. population -- since we have seen that the curve is hollow upward -- moves us down the ladder. Let us try this, trying -1/y first.

Turning to exhibit 6 of chapter 3 we can find

$$-1/5.3 = -.188$$
$$-1/23.2 = -.043$$
$$-1/62.9 = -.016$$

The three points and two slopes are now

$$(1800,-.188)$$
$$(1850,-.043)$$
$$(1890,-.016)$$

$$\frac{-.043 - (-.188)}{1850 - 1800} = \frac{.145}{50} = .0029$$

$$\frac{-.016 - (-.043)}{1890 - 1850} = \frac{.027}{40} = .0007$$

The slope for the first pair of points is now four times that for the second pair. When we used y, that for the second was three times that for the first. We would like equal slopes, so it is natural to try next about halfway from y to -1/y. Thus we are led to try log y. Turning to exhibit 3 of chapter 3, we see that

$$\log 5.3 = .72$$
$$\log 23.2 = 1.37$$
$$\log 62.9 = 1.80$$

so that the three points and two slopes are

$$(1800, 0.72)$$
$$(1850, 1.37)$$
$$(1890, 1.80)$$

$$\frac{1.37-0.72}{1850-1800} = \frac{.65}{50} = .013$$

$$\frac{1.80-1.37}{1890-1850} = \frac{.43}{40} = .011$$

Now the slopes agree with each other rather well. We should now go ahead -- calculating log y and plotting x and log y for either many more or all of the points. Exhibit 6 of chapter 8 has already shown us how well this choice works.

9C. Re-expressing x alone

We now know how to be guided in choosing a better expression for y. What if we wish to leave y alone and re-express x?

If we flip our picture over, interchanging y and x axes, we convert one problem into the other. (We have to look through the back side of the paper after the flip, but

a straight line stays a straight line.) Since it is now the same problem, all the same arguments apply.

This means that if the curve bulges toward large x and we are to re-express x, we ought to move x up the ladder, while if the curve bulges toward small x, we should move x down the ladder.

(* back to the population of the U. S. *)

Exhibit 1 shows three points from the population curve after the interchange of x and y. We see that the bulge is toward larger x (upward on this interchanged plot) so if we are to re-express x we would have to move upward on the x ladder.

Should we try re-expressing x in this situation? No, because we can see that re-expression is unlikely to help us. As amounts, 1810, 1850, and 1890 are very similar; the outer values differing only a few % from the middle one.

Re-expressing a variable that changes by only a few percent rarely gets rid of more than a barely detectable curvature.

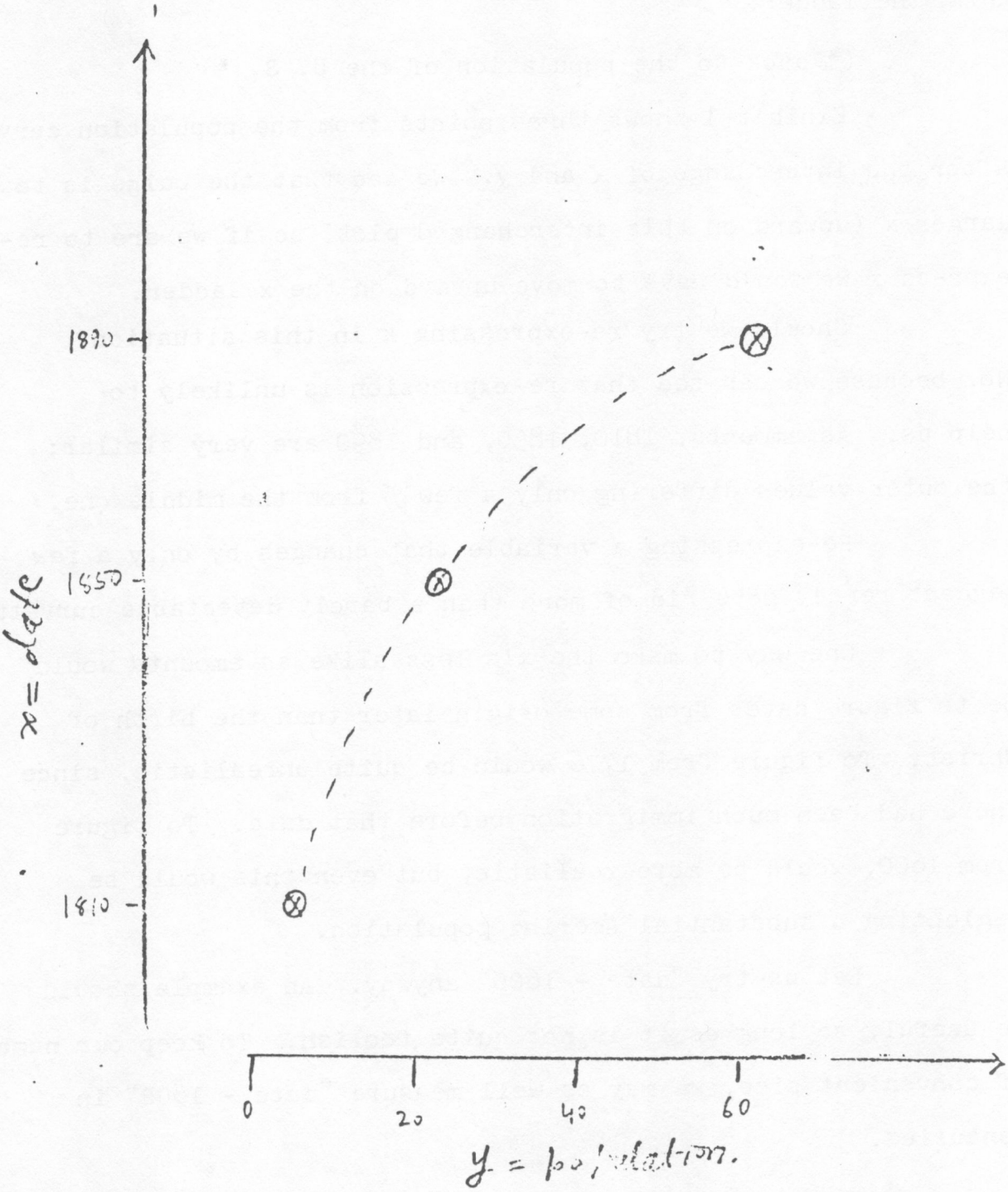
One way to make the x's less alike as amounts would be to figure dates from some origin later than the birth of Christ. To figure from 1776 would be quite unrealistic, since there had been much immigration before that date. To figure from 1600, would be more realistic, but even this would be neglecting a substantial Amerind population.

Let us try "date - 1600" anyway. An example should be useful, so long as it is not quite foolish. To keep our numbers of convenient size, we may as well measure "date - 1600" in centuries.

exhibit 1 of chapter 6

The Early Population of the U.S.

x and y Interchanged



Our new x values for the last 3 points are 2.1, 2.5, and 2.9. They now vary by ±20%, which is much more than before. We have some hope, but we will still not find it easy to take out substantial curvature, such as we face. Since the change from date to $\frac{1}{100}$ (date - 16) involves only multiplication by -- and addition of -- constants, we still need to move up the x ladder.

Let us begin by going to the cube of the new x. We find

$$(2.1)^3 = 9.261$$
$$(2.5)^3 = 15.625$$
$$(2.9)^3 = 24.389$$

The three points are:

(9.3, 7.2)
(15.6, 23.2)
(24.4, 62.9)

and the two slopes are:

$$3.2 = \frac{23.2 - 7.2}{15.6 - 9.3}$$

$$4.5 = \frac{62.9 - 23.2}{24.4 - 15.6}$$

The slope is still larger for the second interval, though only by about 2 to 1 rather than by 3 to 1. We have made progress, but clearly need to go further still.

Trying x^6 gives

(85, 7.2)
(244, 23.2)
(591, 62.9)

and

$$.102 = \frac{23.2 - 7.2}{244 - 85}$$

$$.114 = \frac{62.9 - 23.2}{591 - 244}$$

This is a lot closer, but we are not there yet.

Trying x^8 gives

(378, 7.2)
(1526, 23.2)
(5002, 62.9)

and

$$.0139 = \frac{23.2 - 7.2}{1526 - 378}$$

$$.0114 = \frac{62.9 - 23.2}{4910 - 1526}$$

We seem at last to have gone too far.

Trying x^7 should come quite close. We find

(185, 7.2)
(610, 23.2)
(1725, 62.9)

$$.0368 = \frac{23.2 - 7.2}{610 - 185}$$

$$.0357 = \frac{62.9 - 23.2}{1725 - 610}$$

Agreement is now fairly good.

If we are to re-express

(year - 1600)

we find that our three points suggest the use of either

$$(\text{year} - 1600)^7$$

or, equivalently

$$\left(\frac{\text{year} - 1600}{100}\right)^7.$$

We ought at least look at the results of doing this. Exhibit 2 shows the numbers, exhibit 3 the gross picture, exhibit 4 some differences. As either exhibit 2 or exhibit 4 shows, the deviations of the U.S. population, expressed in millions, from

$$0.72 + 0.036\left(\frac{\text{date} - 1600}{100}\right)^7$$

are less than 150,000 from 1800 to 1830 and from 1870 to 1890, with an "extra" million or two counted in 1840, 1850, and 1860.

* a caveat *

So far either as quality of fit or ability to set residuals in clear view goes, plots such as exhibits 3 and 4 are as good as the plots of the last chapter which related log population to date in years. If our purpose is to examine residuals -- as it so often is -- either plot is effective and useful.

When then should we make a distinction between the two plots? Surely we need to make a distinction when we seek an easily communicated description. That the U.S. population grew about 2.8% per year from 1800 to 1890 is relatively easy both to communicate and to understand. This is a major advantage, since we cannot say the same of the y vs. x^7 relationship. (Both, of course, apply over a limited span of years. Both fail -- faster and faster -- as we move on

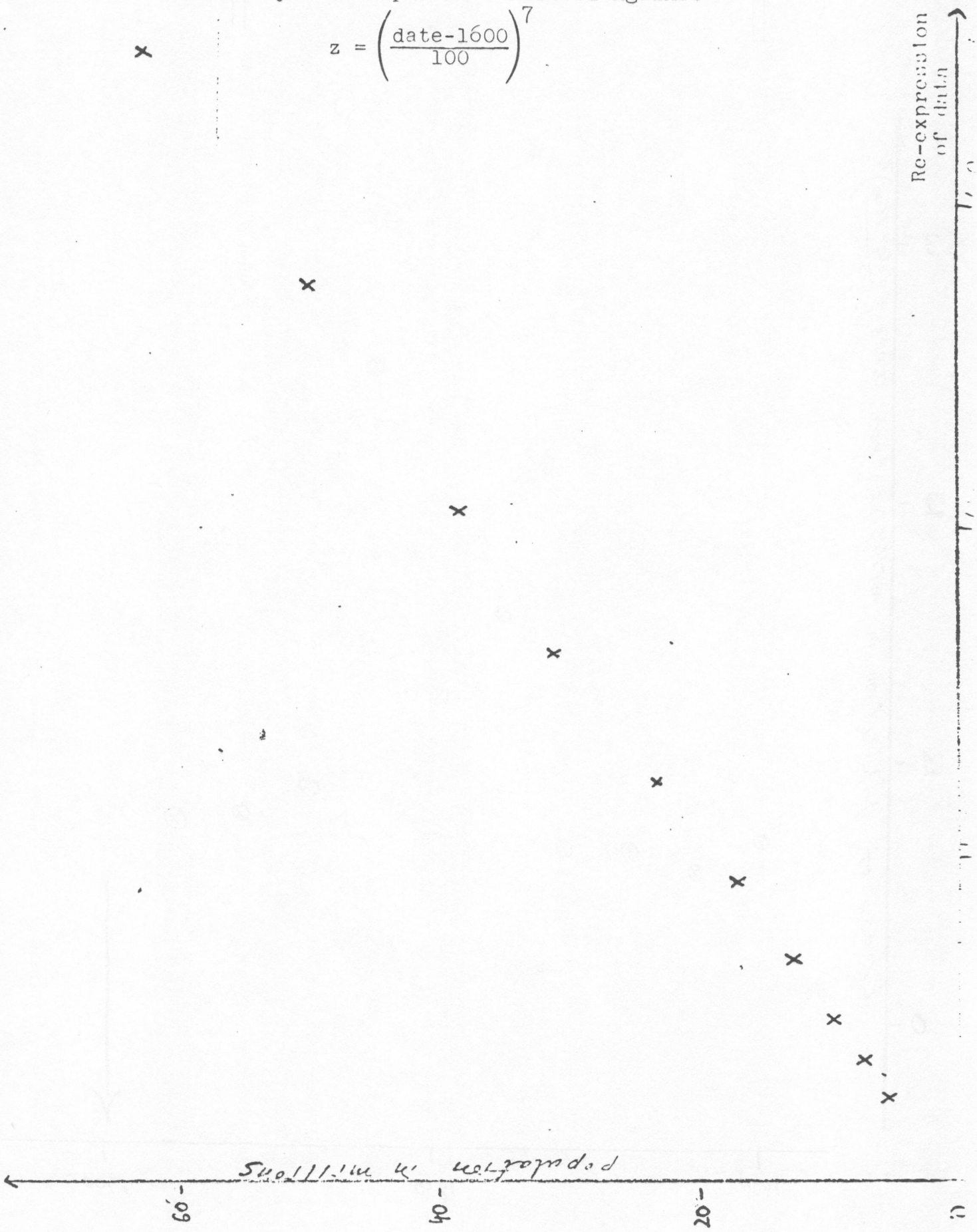
exhibit 2 of chapter 9

The results of re-expressing x in dealing with U.S. population in the Nineteenth Century

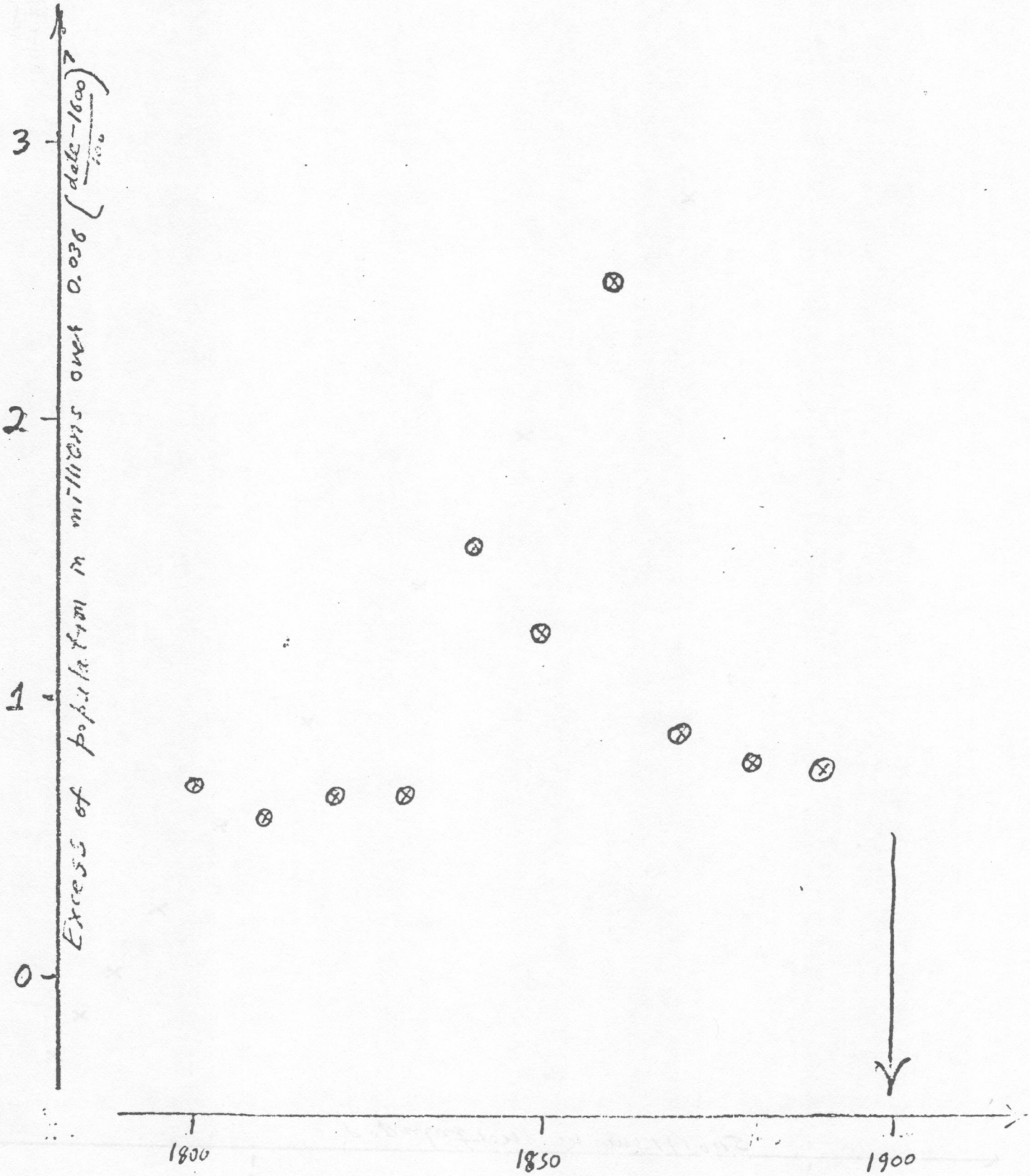
Date	z $\frac{\text{date}-1600}{100}$ [7]	y Population (in millions)	.036z	Diff
1800	128	5.31	4.60	0.71
1810	180	7.24	6.48	0.76
1820	249	9.64	8.98	0.66
1830	340	12.87	12.25	0.82
1840	459	17.07	16.51	1.56
1850	610	23.19	21.97	1.22
1860	804	31.44	28.94	2.50
1870	1047	38.58	37.69	0.87
1880	1372	50.16	49.39	0.77
1890	1725	62.95	62.20	0.75
1900	2187	76.0	78.1	-2.9
1910	2753	92.0	99.1	-7.1
1920	3438	105.7	123.8	-18.1

The Early U.S. Population Plotted Against

$$z = \left(\frac{\text{date}-1600}{100}\right)^7$$



The Result of Flattening exhibit 3



beyond 1900.) For communication, there is no doubt that log y vs. x is a more useful description.

Like any good fit, either the log y vs. x or the y vs. x^7 plot is subject to dangers of overvaluation. We see that each fits closely, though they cannot be exactly alike. Particularly if we have found only one of the two, there is a very natural tendency to convert "a good fit" into "this is how it had to be" or "a basic law of population growth". One example of a close fit, by itself, is far from representing evidence for such strong statements. The fact that we have found two close fits of quite different form emphasizes our need to learn to avoid this sort of jumping at conclusions. Conversely, we can make many good uses of a close fit, whether or not it is "a basic law".

9D. A braking example

Let us next look at an example where re-expressing x seems to be the natural way to make the data more orderly and more describable.

Exhitib 5 shows the speed and distance to stop for 50 cars. Exhibit 6 plots the data. We could fit most of the data points with a straight line. However, a fitted line would give zero stopping distance at a speed between 5 and 10 miles per hour. The one thing we are sure of in this example is that zero stopping distance goes with zero speed, and vice versa. We must face curvature, and try to eliminate it.

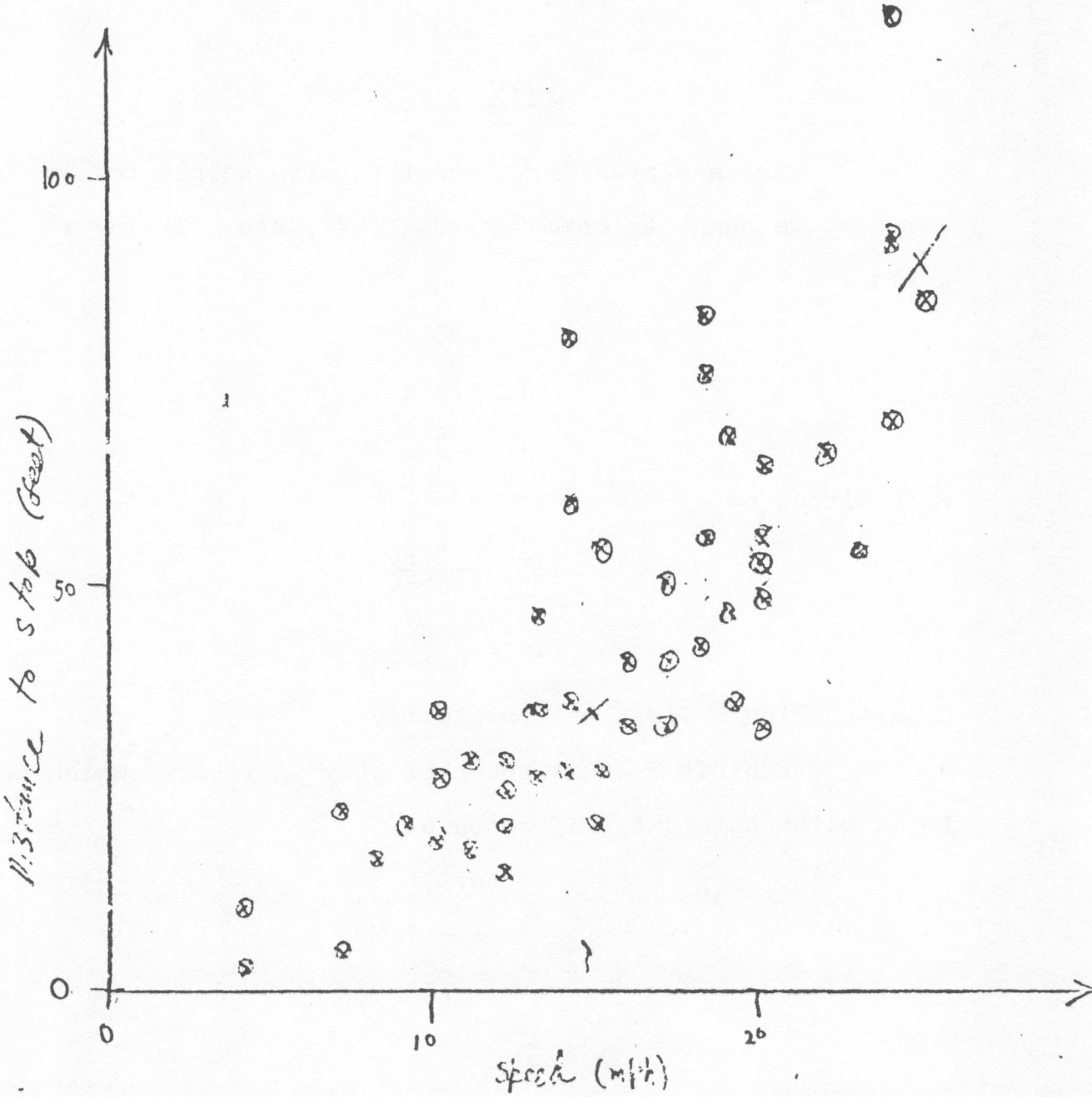
exhibit 5 of chapter 9

Speed and Distance to Stop

Speed, x (mph)	**Distance to stop,** y (in feet)
4	2, 10,
7	4, 22,
8	16,
9	10,
10	18, 26, 34,
11	17, 28,
12	14, 20, 24, 28,
13	26, 34, 34, 46,
14	26, 36, 60, 80,
15	20, 26, 54,
16	32, 40,
17	32, 40, 50
18	42, 56, 76, 84,
19	36, 46, 68,
20	32, 48, 52, 56, 64,
(21)	
22	66,
23	54,
24	70, 92, 93, 120,
25	85,

exhibit 6 of chapter 9

Plot of Exhibit 5's Data



Three reasonable points to take are, then

(0,0)

and the two marked in exhibit 6

(15,35)

and

(25,90)

The two slopes are

$$\frac{35-0}{15-0} = 2.3$$

and

$$\frac{90-35}{25-15} = 5.5$$

To re-express x, since the curve bulges toward large x, we ought to move toward x^2, x^3, etc. Trying x^2 gives

(0,0)

(225,35)

(625,90)

with slopes of

$$.15 = \frac{35-0}{225-0}$$

$$.14 = \frac{90-35}{625-225}$$

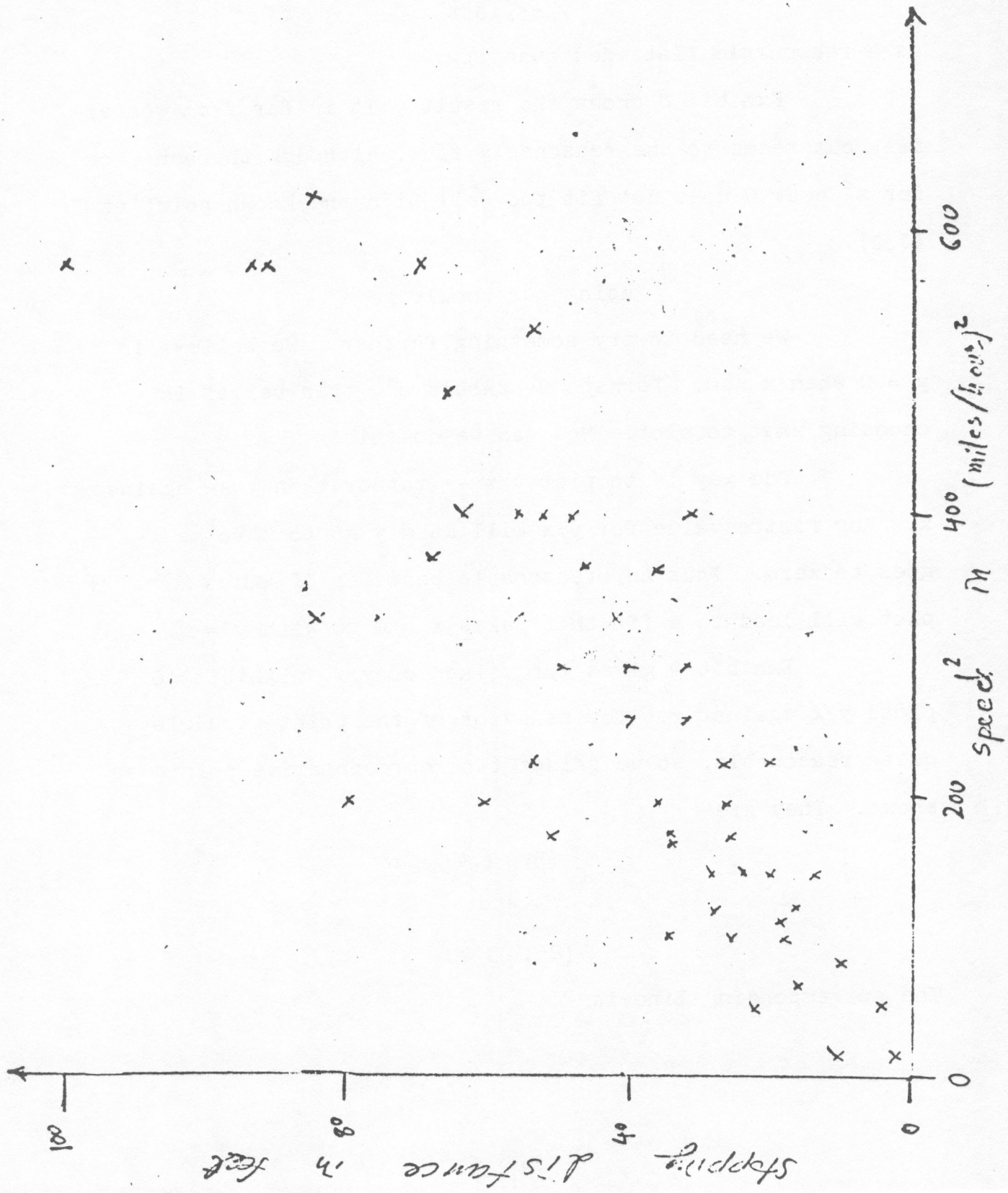
Clearly using x^2 is a reasonable try.

Exhibit 7 shows the plot of y against x^2 which now looks quite straight. If we take

(0,0)

exhibit 7 of chapter 9

Speed^2 and Distance to Stop For 50 Motor Cars



and

$$(600,80)$$

as representative points, we are led to try

$$y - .133x^2$$

as a reasonable flattened quantity.

Exhibit 8 shows the result. It is far from wonderful, but seems to the reasonably flat, although the behavior for x^2 near 0 does not fit too well with the known point at (0,0).

* using our knowledge *

We need to try something further. We believe in $y = 0$ when $x = 0$. Perhaps we should use this belief in choosing what to plot. How can we do this?

One way is to plot y/x -- rather than y -- against x. Any finite value for y/x will make y go to zero as x goes to zero. Thus any reasonable behavior of our y/x vs. x plot will lead to a fit that makes $x = 0$ go with $y = 0$.

Exhibit 9 gives the values of y/x. Exhibit 10 plots y/x against x. The behavior of the point cloud is quite reasonable, so we select two representative points as shown. They are

$$(5,\ 1.4)$$

and

$$(25,\ 3.7)$$

The corresponding line is

exhibit 8 of chapter 9

Flattening of $y - .133x^2$

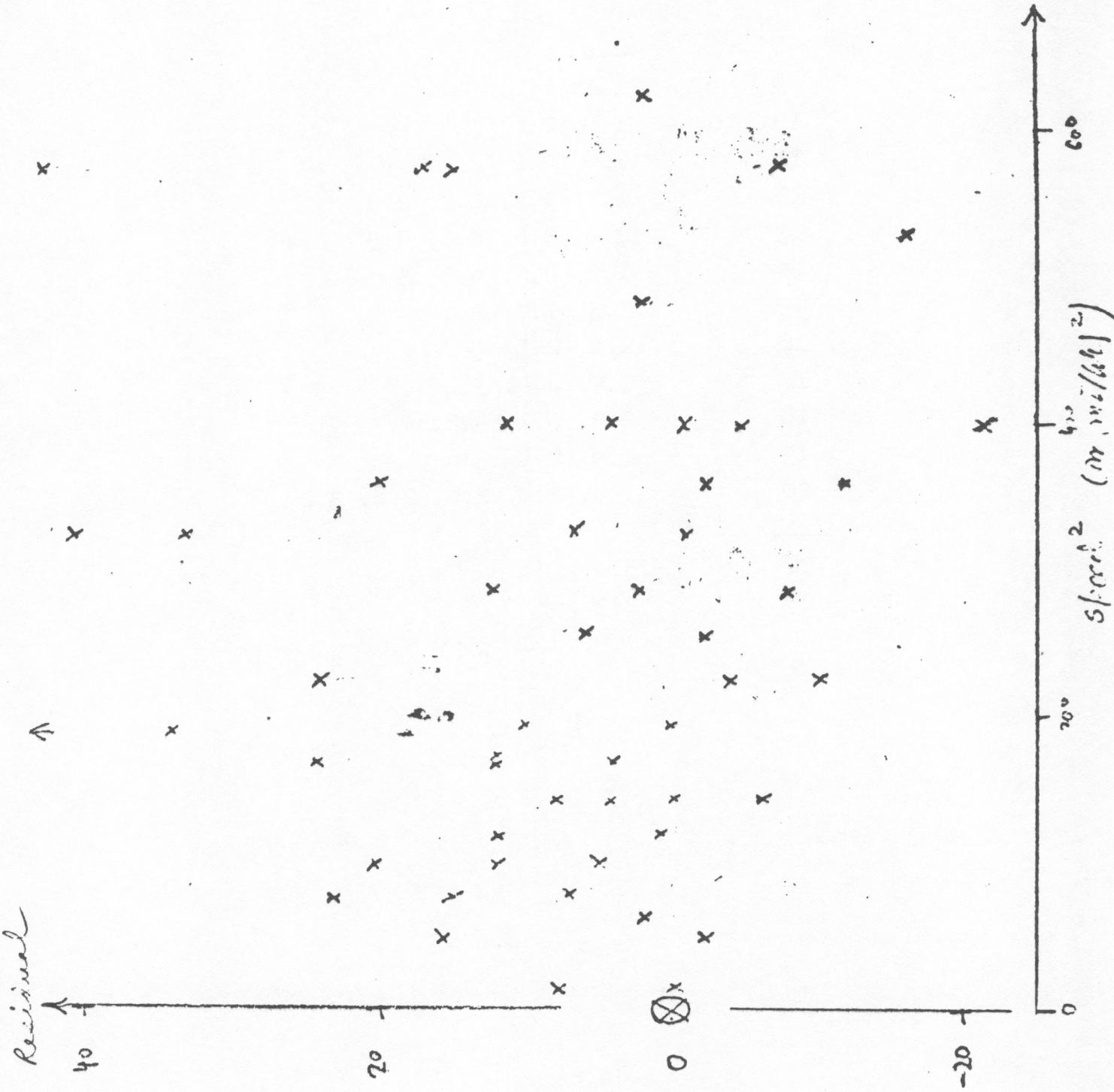
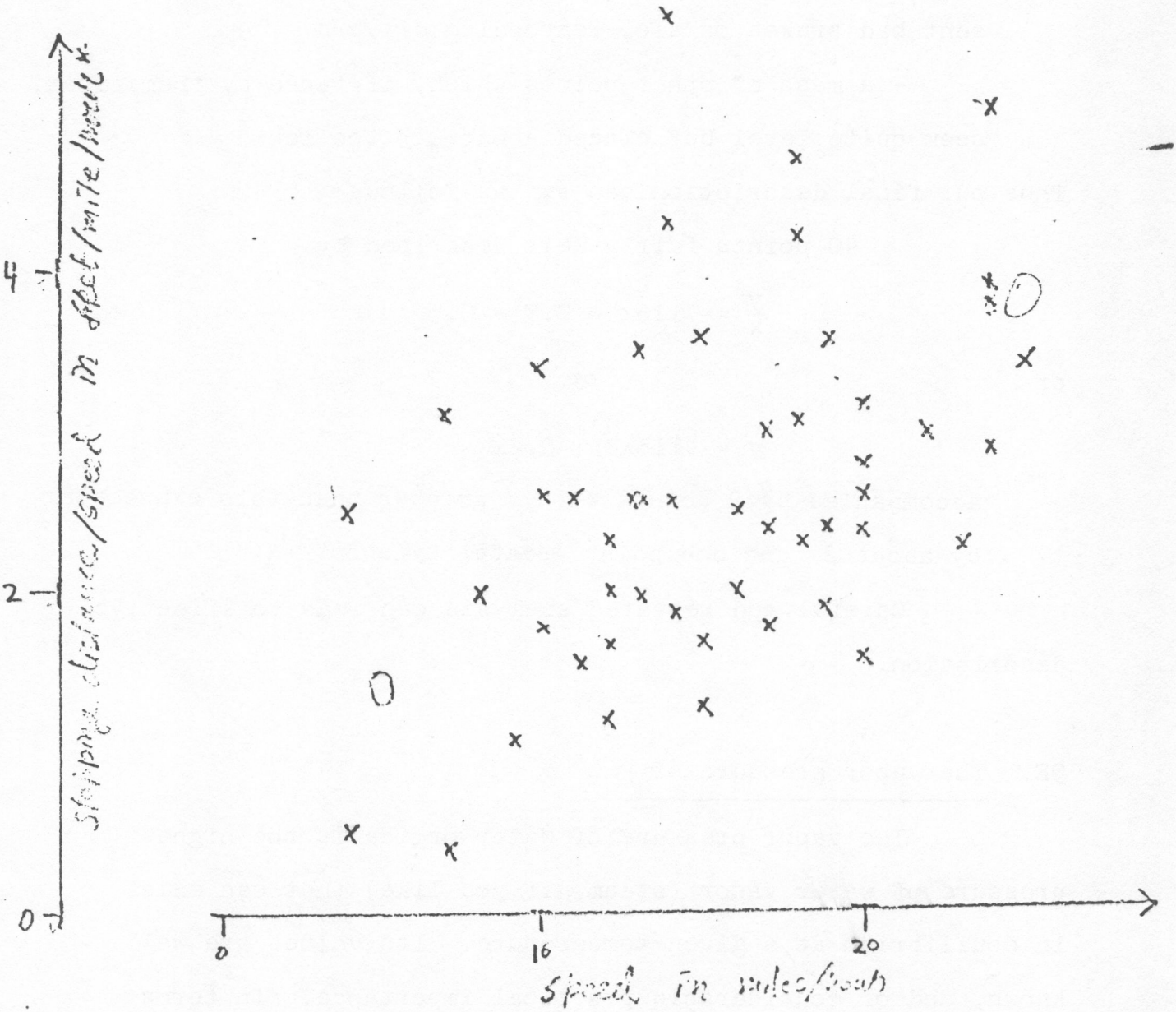


exhibit 9 of chapter 9

The values of y/x

x	y/x
4	0.5, 2.5
7	0.4, 3.1
8	2.0
9	1.1
10	1.8, 2.6, 3.4
11	1.6, 2.6
12	1.2, 1.7, 2.0, 2.3
13	2.0, 2.6, 2.6, 3.5
14	1.9, 2.6, 4.3, 5.
15	1.3, 1.7, 3.6
16	2.0, 2.5
17	1.9, 2.4, 3.0
18	2.3, 3.1, 4.2, 4.7
19	1.9, 2.4, 3.6
20	1.6, 2.4, 2.6, 2.8, 3.2
(21)	
22	3.0
23	2.3
24	2.9, 3.8, 3.9, 5.0
25	3.4

y/x Against x



* 1 ft/mile/hour = $\frac{30}{44}$ second

$$\frac{y}{x} = 0.115x + 0.8$$

The residuals are written down in exhibit 11 and plotted in exhibit 12.

We now see:

- 10 wandering points on the high side -- 9 between 1.1 and 1.9, one very high --- these presumably represent bad brakes or slow-responding drivers.

- a mass of other points which, if taken by themselves, seem quite level but placed about 0.5 too low.

Thus our final description can run as follows:

40 points fairly well described by

$$\frac{y}{x} = .115x + 0.7 - 0.5$$

or or

$$y = .115x^2 + 0.2x$$

accompanied by 9 points with y greater than this expression by about 2x and one point greater by about 4x.

Careful and repeated analysis can lead to effective description.

9E. The vapor pressure of H_2O

The vapor pressure of water or ice is the highest pressure of water vapor (steam, if you like) that can exist in equilibrium at a given temperature. Its values are well known, and of considerable practical importance. In terms

exhibit 11 of chapter 9

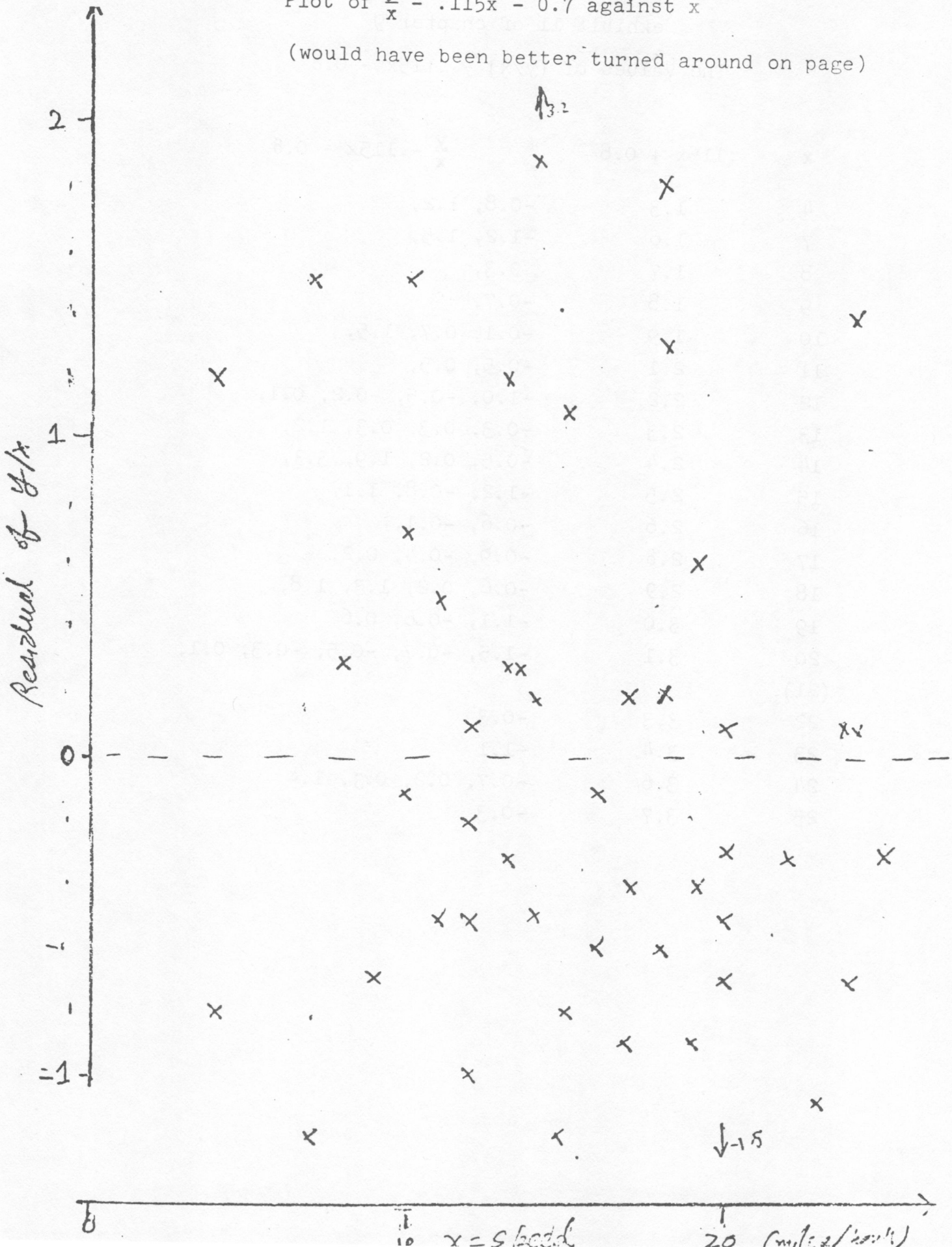
The Values of (y/x) - .115x - 0.8

x	.115x + 0.8	$\frac{y}{x}$ -.115x - 0.8
4	1.3	-0.8, 1.2,
7	1.6	-1.2, 1.5,
8	1.7	0.3,
9	1.8	-0.7,
10	1.9	-0.1, 0.7, 1.5,
11	2.1	-0.5, 0.5,
12	2.2	-1.0, -0.5, -0.2, 0.1,
13	2.3	-0.3, 0.3, 0.3, 1.2,
14	2.4	-0.5, 0.2, 1.9, 3.3,
15	2.5	-1.2, -0.8, 1.1
16	2.6	-0.6, -0.1,
17	2.8	-0.9, -0.4, 0.2,
18	2.9	-0.6, 0.2, 1.3, 1.8,
19	3.0	-1.1, -0.6, 0.6
20	3.1	-1.5, -0.7, -0.5, -0.3, 0.1,
(21)		
22	3.3	-0.3
23	3.4	-1.1
24	3.6	-0.7, 0.2, 0.3, 1.4
25	3.7	-0.3

exhibit 12 of chapter 9

Plot of $\frac{y}{x} - .115x - 0.7$ against x

(would have been better turned around on page)



of temperatures Celsius (once called centigrade, giving 0°C = 32°F and 100°C = 212°F) and pressures in mm of mercury (760 mm is one standard atmosphere) these vapor pressures at different temperatures are given in exhibit 13.

Just a look at the first two columns of this table shows that a direct plot of p against t will only show us a very rapidly rising pressure. (If 139893.20 can be plotted, even 760.00 can hardly be seen.) Accordingly we plan to do something to at least partly straighten out this plot. Taking the logarithm of p looks as if it might help.

Column (3) of exhibit 13 contains the values of log p. This looks conceivably plottable, so we proceed to exhibit 14, which can be looked over, but is far from straight.

Another step needs to be taken. We could try to find some empirical approach, but we can go more directly to our goal if we recall that plotting log pressure against the reciprocal of the absolute temperature is rather common in physical chemistry. (Or if we use the closing example of the next section.) Here the absolute temperature is, in particular, the Celsius temperature increased by about 273.1°C.

Column (4) of exhibit 13 contains values of -1000(1/T), where T = t + 273.1°C. (The factor 1000 has been slipped in to avoid unsightly zeros.) The result of plotting log p against column (4) of exhibit 13 is shown in exhibit 15. At last we have a reasonably good straight line, one worth calculating residuals about.

exhibit 13 of chapter 9

The Vapor Pressure of H_2O and Some Associated Quantities

t(°C)	p(mm Hg)	log p	(4)	(5)
-40	0.105	-.9788	-4.2900	8.6737
-20	0.787	-.1040	-3.9510	8.7858
0	4.5687	.6598	-3.6617	8.8986
20	17.363	1.2396	-3.4118	8.9162
40	54.865	1.7393	-3.1939	8.9256
60	148.88	2.1729	-3.0021	8.9276
80	354.87	2.5501	-2.8321	8.9223
100	760.00	2.8808	-2.6802	8.9112
120	1489.14	3.1729	-2.5439	8.8966
140	2710.92	3.4331	-2.4207	8.8797
160	4636.00	3.6661	-2.3089	8.8611
180	7520.20	3.8762	-2.2070	8.8420
200	11659.16	4.0667	-2.1137	8.8225
220	17395.64	4.2404	-2.0280	8.8034
240	25100.52	4.3997	-1.9489	8.7847
260	35188.00	4.5464	-1.8758	8.7670
280	48104.20	4.6822	-1.8080	8.7502
300	64432.80	4.8091	-1.7449	8.7351
320	84686.80	4.9278	-1.6861	8.7215
340	109592.00	5.0395	-1.6311	8.7095
360	139893.20	5.1458	-1.5795	8.6997

(4) = -1000(1/T), where T = t + 273.1°C

(5) = log p - 2·25(-1000/T)

exhibit 14 of chapter 9

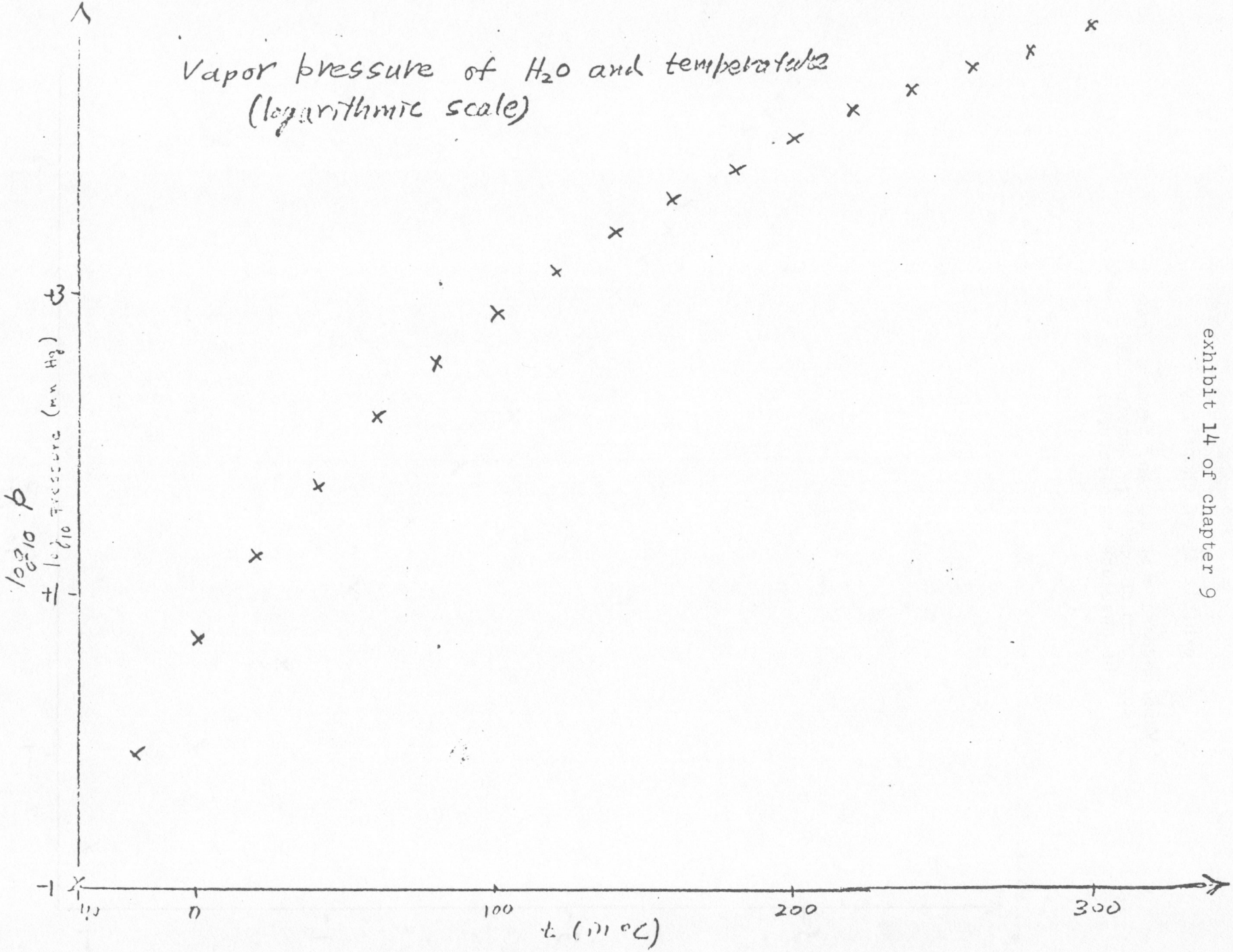
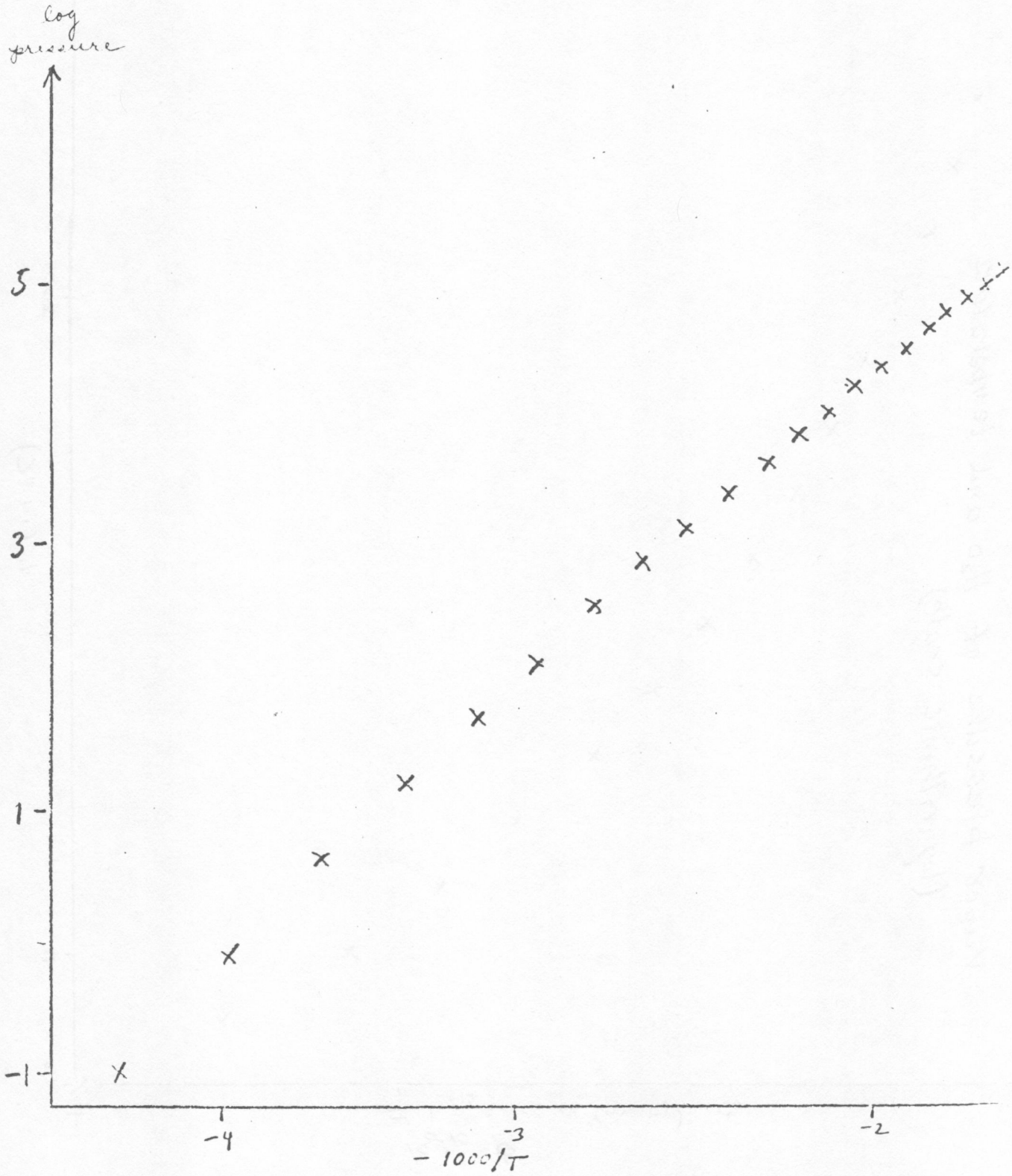


exhibit 15 of chapter 9

Vapor pressure of H_2O and temperature
(logarithm of pressure, reciprocal
of absolute temperature)



Column (5) of exhibit 13 contains values of log p - 2.25 (-1000/T) which, as we see, is reasonably constant. Subtracting 8.8 clearly gives residuals well worth plotting. Exhibit 16 shows the results. The most striking result is the apparent split of our curve into the parts with a corner at the third point from the left. What might this mean?

When we recall that this point comes at 0°C (= 32°F), which is the freezing point of water, we see that a break at this point is quite reasonable. Below 0°C we are dealing with the vapor pressure of solid water (ice), while above this point we are dealing with the vapor pressure of liquid water!

Note that this fact was not forced upon us by our data until we flattened the plot.

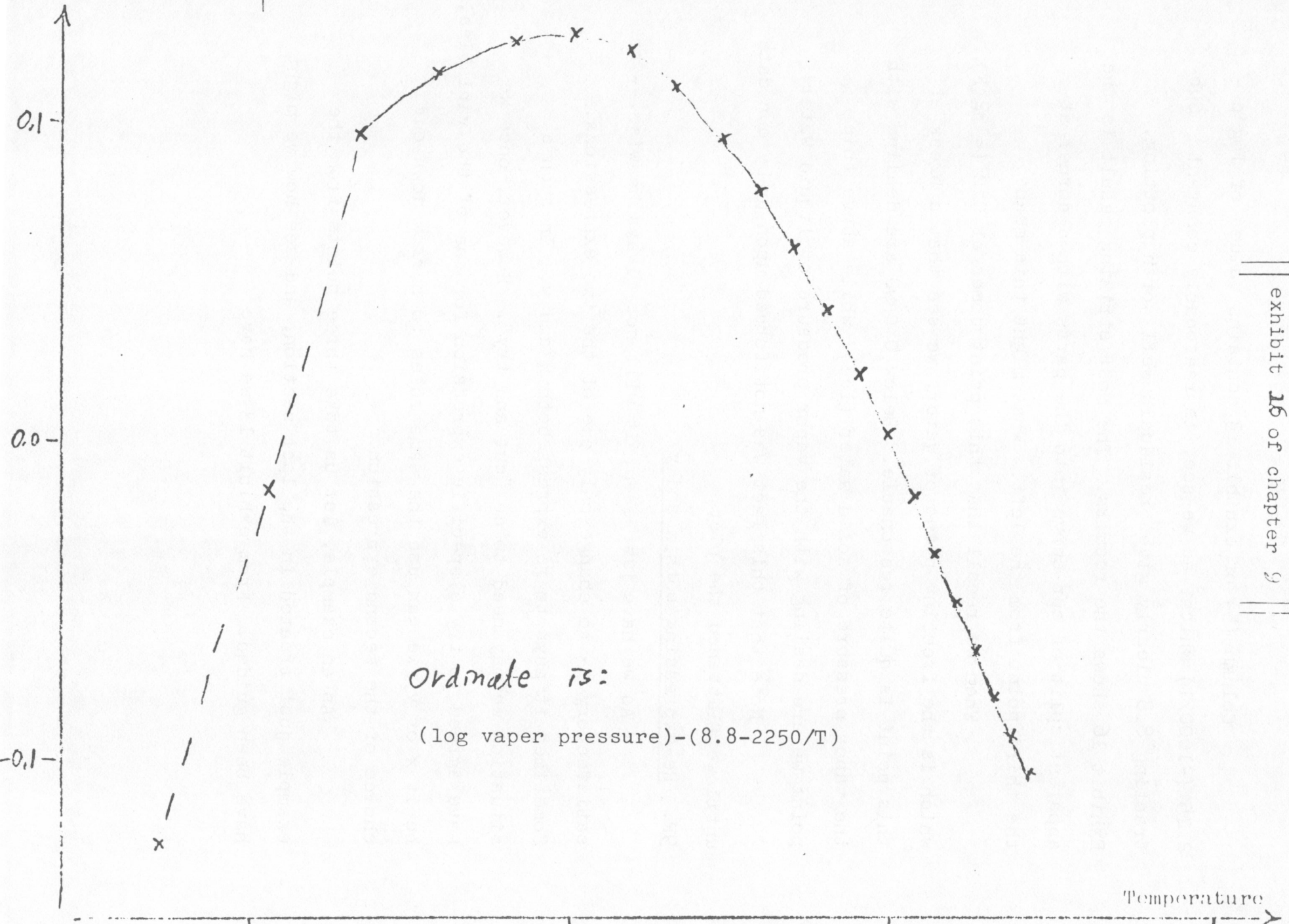
9F. Re-expressing both x and y

As we have just seen, it will not always be wise -- or satisfactory -- to change only one of the two expressions. Sometimes it pays to re-express both x and y. In such a situation we may need to do "cut and try". However, once we have agreed to try a specific expression for one of the variables, be it x or y, we can use the same rules to narrow down our choice of the second expression.

As an example, let us take three points from the example just treated in the last section, and see how we might have been guided. From exhibit 13 we have

exhibit 16 of chapter 9

Residuals of log vapor pressure from a straight line in -1/T
(water)



(0, 4.5687)

(100, 760.00)

(200, 11659.16)

Clearly the middle point is below the line -- and also to the right of the line. If we are to re-express y we should look toward log y and -1/y. Trying logs gives

(0, 0.6598)

(100, 2.8808)

(200, 4.0667)

with slopes of

$$\frac{2.8808-.6598}{100-0} = \frac{2.2210}{100} = .02221$$

and

$$\frac{4.0667-2.8808}{200-100} = \frac{1.1859}{100} = .01186$$

Thus the middle point is now above and to the left of the line. We could try going only as far as $\sqrt{y}$, which is about halfway to log y. If we do this, say by using exhibit 5 of chapter 3, we find

(0, 2.12)

(100, 27.6)

(200, 108)

and we see that $\sqrt{y}$ goes nowwhere nearly far enough. To go on to

$$\sqrt{\sqrt{y}} = y^{\frac{1}{4}} = \sqrt[4]{y}$$

is easy. If we do this, we find

$$(0,\ 1.44)$$

$$(100,\ 5.3)$$

$$(200,\ 10.4)$$

for which the slopes are

$$\frac{5.3-1.44}{100-0} = \frac{3.86}{100} = .0386$$

$$\frac{10.4-5.3}{200-100} = \frac{5.1}{100} = .0510$$

so that we see that we are not yet far enough. Those who love to repeat square roots will now try

$$\sqrt{\sqrt[4]{y}} = y^{\frac{1}{8}} = \sqrt[8]{y}$$

finding

$$(0,\ 1.20)$$

$$(100,\ 2.32)$$

$$(200,\ 3.20)$$

for which the slopes are

$$\frac{2.32-1.20}{100-0} = \frac{1.12}{100} = .0112$$

$$\frac{3.20-2.32}{200-100} = \frac{.88}{100} = .0088$$

This time we have gone too far. As a result, if we are to change only y, we should try something between $y^{\frac{1}{4}}$ and $y^{\frac{1}{8}}$, perhaps $y^{1/6}$ or $y^{1/7}$.

*another try *

This is, however, not our only reasonable choice. Going to log y did much to straighten out our three points. True it went a little too far, but perhaps we could do something by keeping log y and re-expressing x. Before we do

this, however, we should stop and think for a moment. As we have written it, x is temperature in °C (freezing water = 0°C, boiling water = 100°C). If we are to re-express simply, we ought not to tie our zero to a property of so special a substance as water. Rather we should figure our temperatures from the so-called absolute zero, which is just below -273°C.

Our intermediate starting point then, should be

(273.1, 0.6598)

(373.1, 2.8808)

(473.1, 4.0667)

with slopes of .02221 and .01186, and a middle point above and to the left of the line. If we are to re-express x, we should move toward log x or -1/x. Let us try log x. We find:

(2.44, 0.6598)

(2.57, 2.8808)

(2.67, 4.0667)

$$\frac{2.8808-0.6598}{2.57-2.44} = \frac{2.2210}{.13} = 17.1$$

$$\frac{4.0667-2.8808}{2.67-2.57} = \frac{1.1859}{.10} = 11.8$$

Using logs has helped, but not enough. Let us try -1/x, for which we have

(-.0037, 0.6598)

(-.00270, 2.8808)

(-.00212, 4.0667)

$$\frac{2.8808 - 0.6598}{-.00270 - (-.0037)} = \frac{2.221}{.0010} = 2221$$

$$\frac{4.0667 - 2.8808}{-.00212 - (-.00270)} = \frac{1.186}{.00058} = 2045$$

for which the slopes agree better than any other combination so far tried.

Accordingly

$$\log y$$

and

$$-1/x$$

where x is absolute temperature, seems to be a good choice. As we know from the last section, it proves to be one.

9G. Wise change of origin as a preliminary

We have already seen one example where a change of origin was part of a sensible approach to straightening: the early population of the U.S. as a function of time, where a time origin at the birth of Christ -- 1600 years before the beginning of European immigration to North America -- was far from sensible -- once we thought about it! Why 1600 and not 600 or 2600?

The same sort of need to think about a sensible origin arises in other problems, sometimes with regard to x and sometimes with regard to y. In one broad class of cases, the question arises because the amount we measure is the sum of contributions, some of which change slowly if at all, while the remaining contribution changes in such a way that re-expression could flatten the plot if this contribution could be measured separately.

The natural approach to this sort of data is to form

observation - background

where "background" is a constant chosen to allow for the slowly changing components. We are not likely to be able to pick the value for "background" either on the basis of general insight or on the basis of doubtfully-related historical data. We expect to learn about it from the same set of data that we are trying to flatten. We are likely to approach choosing a plausible value for "background" by trying various values for it and seeing which one leads to data that can be more thoroughly flattened by further re-expression.

Observation of radioactive decay yields many problems of this sort. Any single kind of radioactive atom decays on a steady percentage basis -- so many % is gone every so many days, years, or millenia. Many processes of isolating -- or making -- some one kind of radioactive atom also isolate -- or make -- one or more other kinds. If these other kinds decay slowly, their presence can often be adequately allowed for by a constant background.

* a radioactive decay example *

In 1905, the study of radioactive substances was in its infancy. Meyer and von Schweidler reported the relative activities for an experimental object set out in exhibit 17. If we plot activity against time, the result is exhibit 18, which is far from being straight. It is again

exhibit 17 of chapter 9

The Decay of Radioactivity and Associated Quantities

time (days)	activity (relative)	(3)	(4)	(5)	(6)	(7)
0.2	36.0	1.556	1.544	1.538	1.551	1.545
2.2	26.0	1.415	1.398	1.389	1.471	1.466
4.0	23.1	1.364	1.344	1.334	1.473	1.474
5	18.9	1.276	1.253	1.241	1.418	1.416
6	17.8	1.250	1.225	1.212	1.421	1.422
8	14.7	1.167	1.137	1.121	1.401	1.401
11	13.4	1.127	1.093	1.076	1.456	1.461
12	11.3	1.053	1.013	.991	1.409	1.411
15	8.5	.929	.875	.845	1.370	1.370
18	5.9	.771	.690	.643	1.286	1.273
26	5.0	.699	.602	.544	1.460	1.454
33	3.4	.534	.380	.279	1.469	1.434
39	2.4	.380	.146	-.046	1.431	1.319
45	2.1	.322	.041	-.222	1.526	1.353

(3) = log(activity)

(4) = log(activity - 1.0)

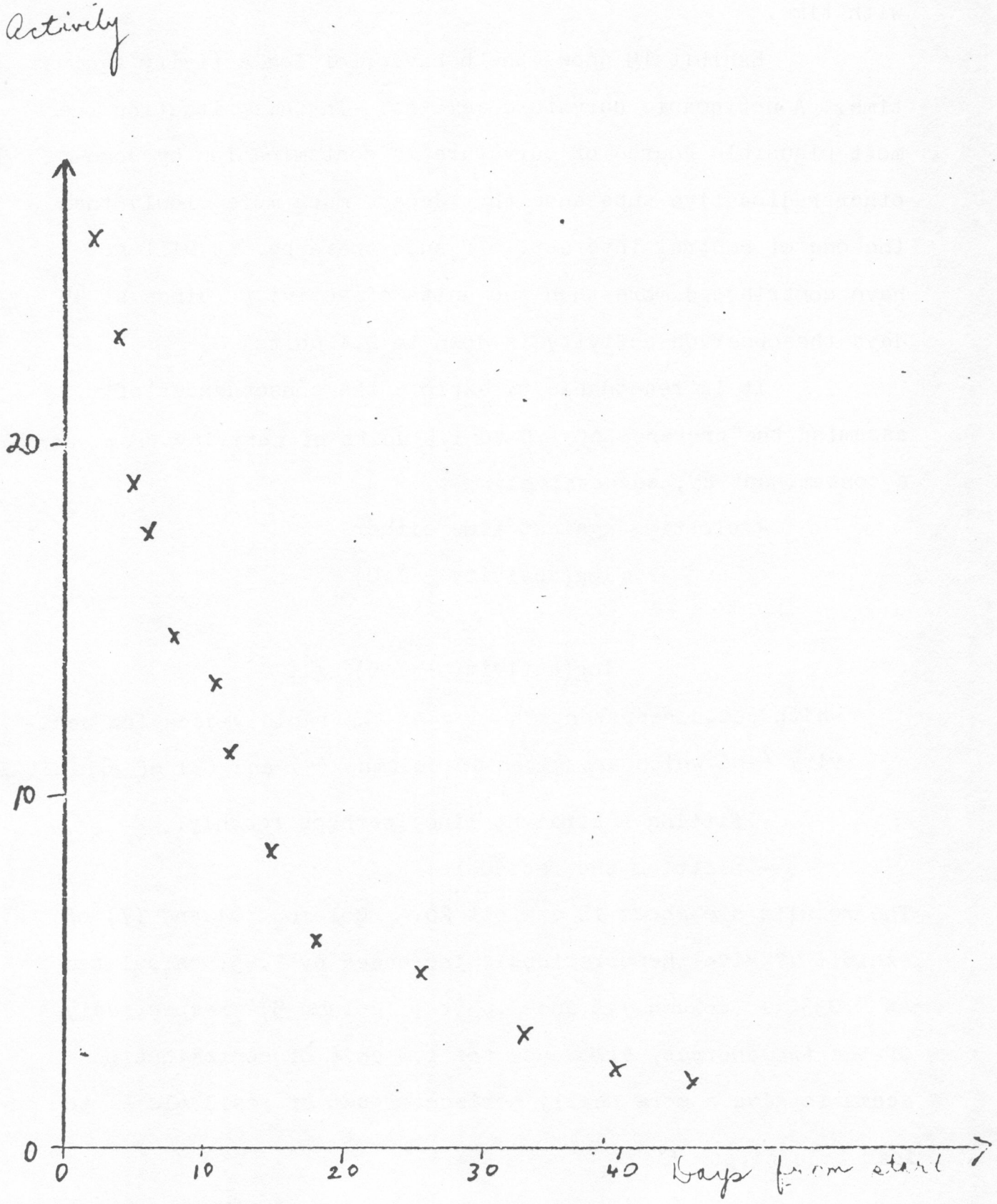
(5) = log(activity - 1.5)

(6) = .033t + (column 4)

(7) = .035t + (column 5)

Source: S. Meyer and E. von Schweidler 1905. Sitzungsberichte der Akademie der Wissenschaften zu Wien, Mathematisch-Naturwissenschaftliche Classe, p. 1202 (Table 5).

exhibit 18 of chapter 9
An Example of Radioactive Decay



natural to take logarithms, especially since, in simple radioactive decay, the logarithm of activity should decrease linearly with time.

Exhibit 19 shows the behavior of log activity against time. A noticeable curvature remains. In this situation the most plausible source of curvature is contamination by some other radioactive substance that decays much more slowly than the one of central interest. If such there be, it will not have contributed more than two units of activity (since at 45 days the observed activity is down to 2.1 units).

It is reasonable to explore the consequences of assuming the presence of 1.0 to 1.5 units of activity from such a contaminant by, successively:

- plotting against time either

  log(activity - 1.0)

  or

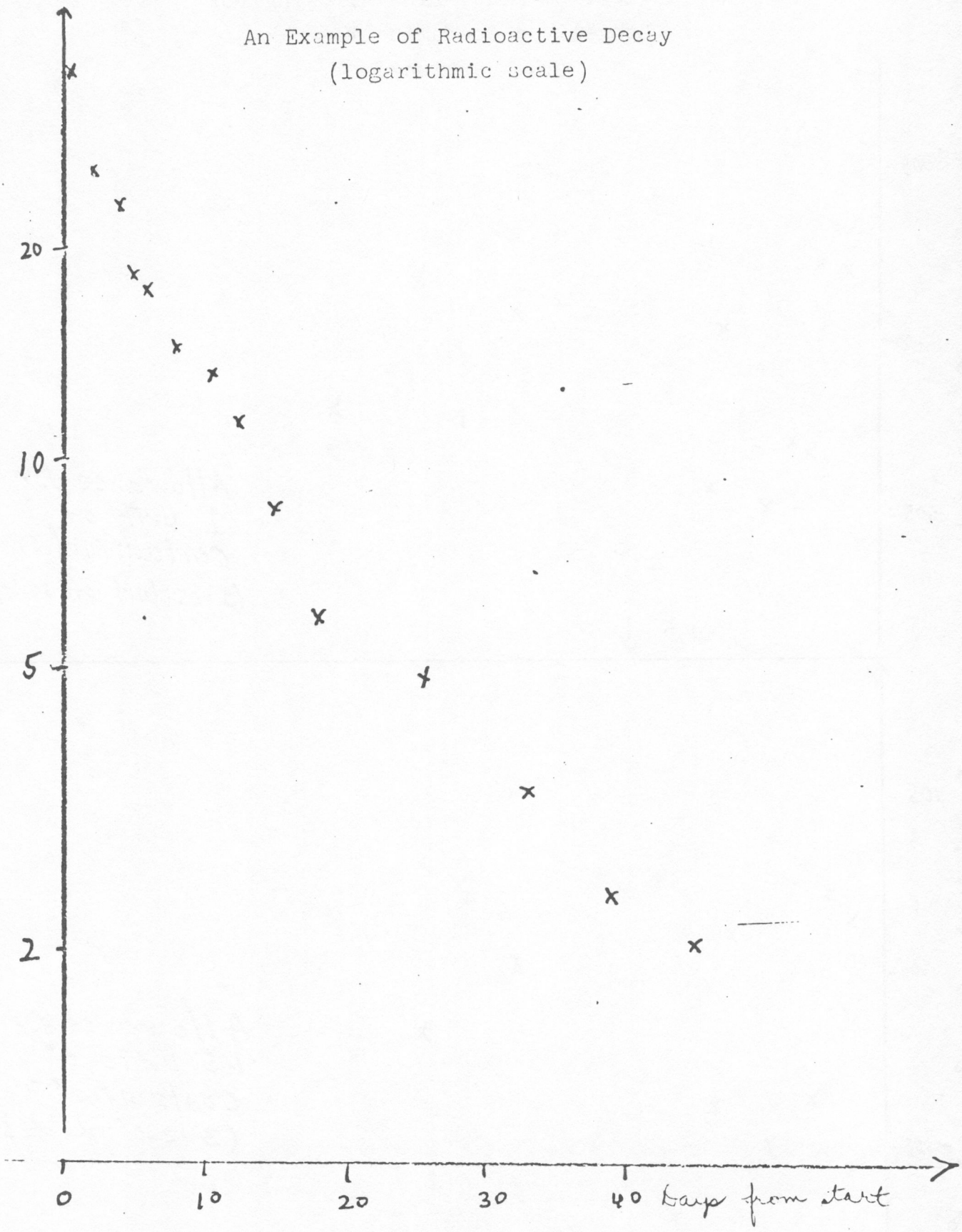
  log(activity - 1.5)

  which could represent the log of the rapidly-decaying activity (and which are given in columns (4) and (5) of exhibit 17,

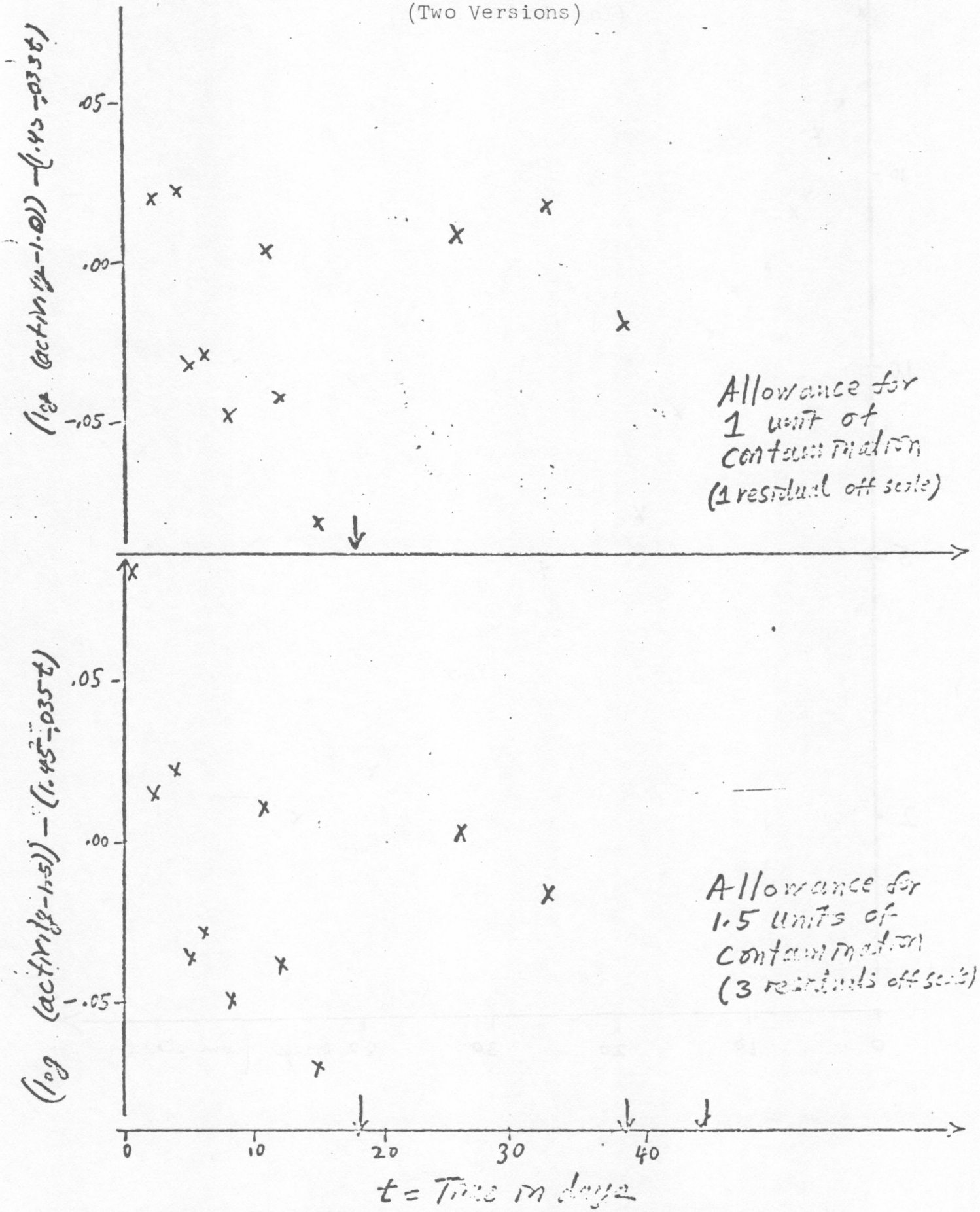
- fitting a straight line, perhaps roughly.
- plotting the residuals.

The results are shown in exhibit **20**. (Columns (6) and (7) of exhibit 17 give these residuals increased by 1.45, calculated as ".033t + (column 4)" and ".035t + (column 5)" respectively.) Of the two choices, allowance for 1.0 unit of contamination seems to give a more nearly horizontal set of residuals -- to lead to a closer fit.

An Example of Radioactive Decay
(logarithmic scale)



Residuals from Decay Line Allowing for
Long-lived Contamination
(Two Versions)



(Perhaps the residuals for 1.0 do trend upward a little. The reader may wish to try 1.1 or 1.2.)

*****

9H. How far have we come?

This chapter has been devoted to guidance about how to re-express x and y with a view to straightening a plot. This question arises most simply when one or both of x and y is an amount -- or, really just a special case of an amoung,a large count. The natural re-expressions are by powers, roots, and log

The basic rule of thumb is simple:

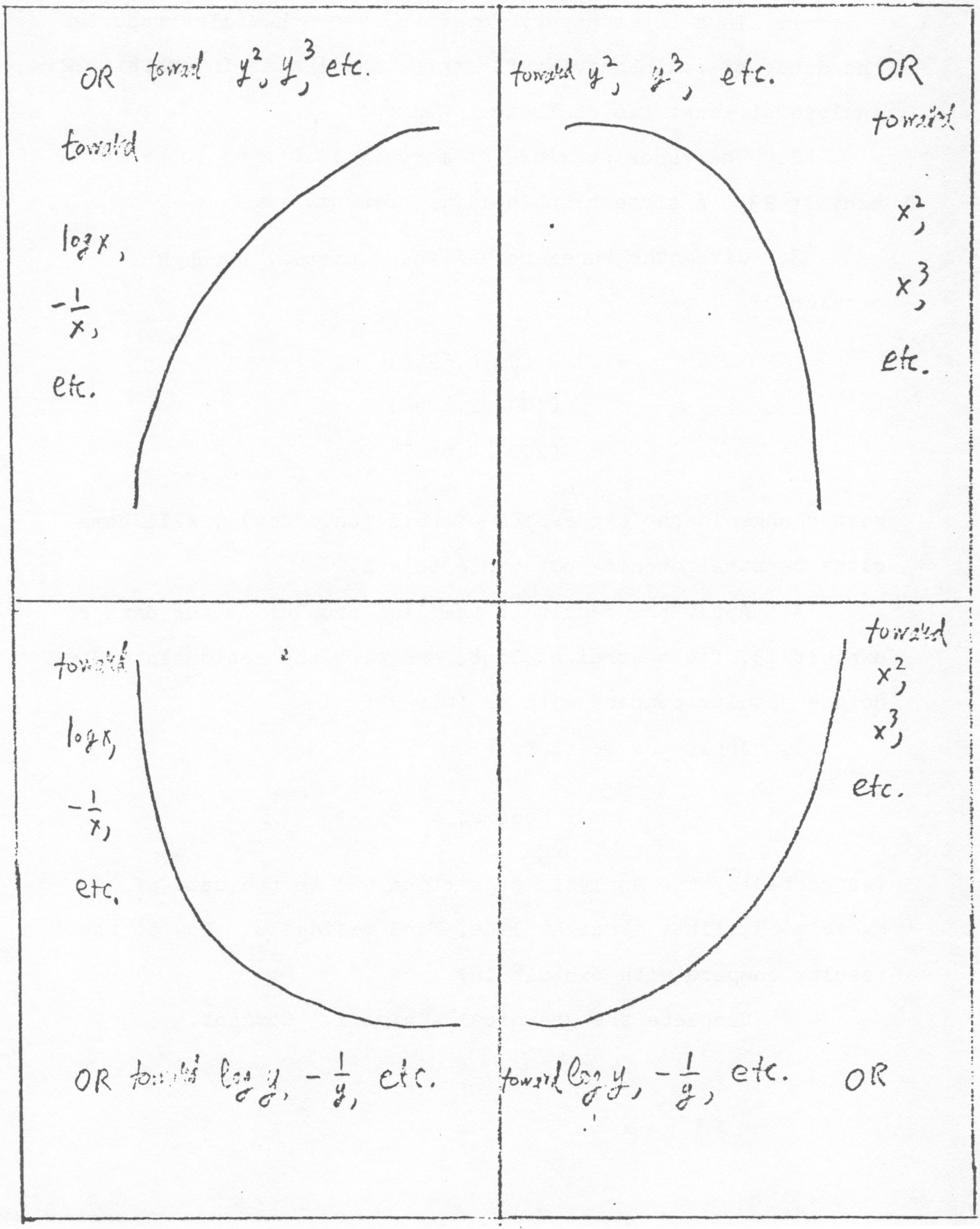
- move on the ladder of expressions in the direction in which the curve bulges.

Exhibit 21 shows the four possible cases, and the natural steps along both x- and y-ladders.

To keep computation down, we routinely begin with three well-chosen points and only try new expressions on more of the data when the three points respond well to them.

In certain cases, a fresh choice of origin -- before going to powers, roots or logs -- is valuable. Sometimes the new choice is a matter of common sense, sometimes of how flat the final result proves to be.

How to Move for Each Variable Separately the Four Cases



*****

9P. Problems

1. In a later paper, Meyer and von Schweidler reported the decay of radioactivity for three samples as in exhibit 22. Analyze at least two of these. Comment.

2. The vapor pressure of mercury is stated to be as in exhibit 23. Analyze graphically. Comment.

3. Given the three points found partway through section 9F

(0, 0.6598)

(100, 2.8808)

(200, 4.0667)

what change in the expression of this (once new) y will come close to straightening out these points.

4. Apply the result of the last problem to the data of exhibit 13, fit a straight line, and plot the residuals. How do the results compare with exhibit 16?

5. Apply the re-expression

$$y \text{ becomes } y^{1/6}$$

(suggested by the analysis of section 9E) to the data of exhibit 13, fit a straight line, find residuals. How do the results compare with exhibit 16?

Complete the graphical analysis. Comment.

exhibit 22 of chapter 9

Three Examples of Radioactive Decay

Days*	Activity**	Days*	Activity**	Days*	Activity**
0.8	6.70	0.8	2.82	0.8	2.05
2.8	6.40	2.8	2.34	1.0	2.03
6.9	5.70	4.8	1.90	1.8	1.79
8.9	5.10	6.8	1.80	2.1	1.77
13.1	4.30	11.8	1.34	3.9	1.54
15.2	4.00	13.9	1.24	5.9	1.35
16.8	3.95	16.8	1.03	7.1	1.29
20.1	3.40	19.8	1.00	9.1	1.23
20.8	3.40	23.8	0.80	12.1	1.01
20.9	3.20	31.8	0.55	12.9	0.96
21.9	3.20	32.5	0.11	16.9	0.82
31.1	2.42			19.9	0.68
36.8	2.30			22.8	0.59
43.8	2.11			25.9	0.51
49.8	2.00			33.8	0.405
53.8	1.99			44.1	0.32
58.8	1.98				
65.9	1.90				
73.8	1.80				
87.1	1.65				

* Since start.

** In volts/minute.

Source: S. Meyer and E. von Schweidler 1907. Untersuchungen über radioaktive substanzen. (VIII Mitteilung) Über ein radioaktives Produkt aus dem Aktinium. Sber Ak Wiss Wien Math-Nat Class 116, IIA1, pp. 315-322 (especially 316-317).

exhibit 23 of chapter 9

Vapor Pressure of Mercury

Temperature (°C)	Pressure (mm Hg)
0	0.0004
20	0.0013
40	0.006
60	0.03
80	0.09
100	0.28
120	0.8
140	1.85
160	4.4
180	9.2
200	18.3
220	33.7
240	59.
260	98.
280	156.
300	246.
320	371.
340	548.
360	790.

(Source: Gebhardt, Hertz, Regnault, van der Plaats, and others.)

6. Rates of mortality from breast cancer in different latitudes. (50;1025,513) translates as: "In latitude 50°N, the mortality index for breast cancer is 102.5 and the mean annual temperature is 51.3. Data-sets (16): (50,1025,513), (51,1045,499), (52,1004,400), (53,959,492), (54,870,485), (55,950,478), (56,886,473), (57,892,451), (58,789,463), (59,846,421), (60,817,442), (61,722,435), (62,651,423), (63,681,402), (69,673,318), (70,525,340). Source: A. J. Lea 1965. New observations on distribution of neoplasms of female breast in certain European countries. British Medical Journal 1 (for 1955): 486-490. Table II on page 489. Problem: Can we straighten this plot?

7. Plasticity of wool: slow stretching of a single fiber. (1,321) translates as: "After 1 minute under load, a fiber 53.3 microns in average diameter (coefficient of variation of diameter = 5.4%) had stretched 32.1% beyond its unloaded length." Data-sets (34): (1,321), (3,330), (5,334), (8,337), (16,342), (32,348), (50,352), (110,361), (240,377), (440,394), (740,413), (1310,442), (1460,449), (1630,458), (1900,469), (2090,478), (2760,499), (2950,505), (3080,509), (3460,519), (4280,540), (4970,556), (5720,572), (6000,579), (6320,586), (7120,600), (7360,604), (7540,607), (8520,623), (9020,629), (9230,633), (9950,643), (10260,647), (10680,654). (The last data set is for 10,680 minutes under load and 65.4%

stretch.) Source: O. Ripa and J. B. Speakman 1951. The plasticity of wool. Textile Research Journal 21: 215-221. Table I on page 217. Problem: How can we straighten the plot?

8. Amount of interstitial space in young chickens. ("Interstitial space" is a volume defined by the amount of thiocyanate ion taken up outside the blood within 10 minutes after injection and measured by its ratio to blood volume.) (1,52) translates as: "For chickens 1 week old, the interstitial space accounted for 52% of body weight." Data-sets (8): (1,52), (2,42), (3,39), (4,38), (6,37), (8,36), (16,25), (32,22). Source: W. Medway and M. R. Kare 1959. Thiocyanate space in growing fowl. Amer. J. of Physiology 196: 873-875. Table I on page 874. Problem: How can we straighten the plot?

9. A more detailed look at interstital volume. (1,55,52) translates into: "At age 1 week, the average body weight of 6 chickens was 55 grams, of which the interstitial space was 52%". Data-sets (8): (1,55,52), (2,108,42), (3,175,39), (4,242,38), (6,372,37), (8,527,36), (16,1137,25), (32,1760,22). Source: As for problem 8. Problem: How can we straighten the plot? How does the straightness here compare with that for problem 8? What expression of amount (not %) of interstitial space in terms of weight corresponds to our fit?

10. Vapor pressure of a boron analog of mesitylene. (130,29) translates as: "At a temperature of 13.0°C, the vapor pressure of B-trimethylborazole was 2.9 millimeters of mercury." Data-sets (13): (130,29), (195,51), (255,85), (272,103), (318,146), (384,213), (457,305), (561,514), (644, 745), (714,1002), (805,1437), (857,1769), (915,2169). Source: E. Wiberg, K. Hertwig, and A. Bolz 1948. Zur kemmtnis der beiden symmetrischen Trimethyl-borazole ("anorganisches Mesitylen"), Zeitschrift für Anorganische Chemie 256: 177-216. Table on page 191. Problem: How can we straighten this plot? Check your answer!

11. Effects of small amounts of biotin on the mobility of a micro organism. (5E-7,1354) translates as: "At a biotin concentration of 5×10^{-7}, the mobility of Lactobacillus casei was 1.354 units." Data-sets (7): (0,1415), (5E-7,1354), (1E-6,1311), (5E-6,1230), (1E-5,1234), (5E-5,1181), (1E-4,1188). Source: V. R. Williams and H. B. Williams 1949. Surface activity of biotin. Journal of Biological Chemistry 177: 745-750. Problem: How can we straighten the plot?

12. Electric current due to heating aluminum phosphate. (880,1) translates as: "At a temperature of 880°C -- an absolute temperature of 880 + 273 = 1153°K -- positive electrification produced a current of 1 unit, where 1 unit = 2×10^{-9} amperes." Data-sets for run A (8): (880,1), (950,4), (970,7), (995,15), (1030,35), (1055,49), (1095,103), (1110,126).

Data-sets for run B (9): (1036,1), (1088,3), (1135,5), (1160,8), (1195,15), (1230,34), (1245,35), (1295,74), (1330,168). Source: A. E. Garrett 1910. Positive electrification due to heating aluminium phosphate. (London, Edinburgh and Dublin) Philosophical Magazine 20: 571-591. Table on page 581. Problem: How can we straighten the plots?

13. Growth-promoting effect of a purine for a deficient strain of red bread-mold. (0,112) translates as: "When 0 moles of a guanine per mole of adenine were included with 0.1 milligrams of adenine in 25 milliliters of basal medium, the dry weight of mycelium produced by a purine-deficient strain of Neurospora was 11.2 milligrams." Data-sets (9): (0,112), (.25,135), (.59,152), (.75,185), (1,196), (1.5,203), (2,203), (2.5,243), (3,244). Source: J. L. Fairley, Jr., and H. S. Loring 1949. Growth-promoting activities of guanine, guanosine, guanylic acid, and xanthine for a purine-deficient strain of Neurospora, Journal of Biological Chemistry 177: 451-453. Table I on page 453. Problem: How can we straighten the plot?

14. Measurement of a certain impurity in DDT; change of scale factor with temperature. (21,248) translates as "At 21°C, the rate of crystallization (in microns per 5 minutes) is 24.8 times the log of the % of this particular impurity. Data-sets (14): (21,248), (22,308), (23,388), (24,465),

(25,569), (26,678), (27,806), (28,959), (29,114), (30,139), (31,168), (32,202), (33,236), (34,270). Source: W. McCrone, A. Smedal, V. Gilpin 1946. Determination of 2,2, bis-p-chlorophenyl -1,1,1- trichloroethane in technical DDT. A microscopical method. Industrial and Engineering Chemistry 18: 578-582. Table IV on page 582. Problem: How can this plot be straightened.

15. Demand deposits in post office savings accounts in Switzerland. (37,458) translates as: "In 1937, there were 458 million francs in post office savings accounts." Data-sets (29): (37,458), (38,498), (39,523), (40,643), (41,701), (42,787), (43,839), (44,927), (45,1001), (46,1079), (47,1007), (49,1033), (49,1090), (50,1125), (51,1212), (52,1248), (53, 1334), (54,1393), (55,1443), (56,1720), (57,1720), (58,1896), (59,2050), (60,2268), (61,2643), (62,3140), (63,3353), (64, 3513), (65,3810). Source: Swiss Statistical Abstract, issued November 1969 by the Swiss Credit Bank. (Title also in French and German.) Tables at pages 24 and 25. Problem: How can we best straighten a plot for this data? Make the fit and plot the residuals. Summarize all results.

16. Revenue passenger miles on U.S. passenger airlines. (37,412) translates as: "In 1937, there were 412,000 revenue passenger miles on U.S. domestic scheduled airlines." Data-sets (24): (37,412), (38,480), (39,683), (40,1052), (41,1385),

(42,1418), (43,1634), (44,2178), (45,3362), (46,5948), (47,6109), (48,5981), (49,6753), (50,8003), (51,10566), (52,12528), (53,14760), (54,16769), (55,19819), (56,22362), (57,25340), (58,25343), (59,29269), (60,30514). (The last translates as: "In 1960, 30,514,000 revenue passenger miles.) Source: Robert G. Brown 1963. Smoothing, Forecasting and Prediction of Discrete Time Series, 1963, Prentice-Hall. Table C.7 on page 427. His source: F. A. A. Statistical Handbook of Aviation. Problem: How can this plot be straightened? Plot residuals from a well-chosen straight-line fit.

17. Find, from other sources, 2 batches of (x,y) data-sets that interest you and deserve to have their plots straightened by the methods of this chapter. Straighten the plots. Plot the residuals.